Optical character recognition in the US Government

E. R. Lannon

Input to a computer is still one of the slowest elements in automatic data processing. If we compare the internal speed of the computer system scheduled to be installed at the National Institutes of Health of my Department we find that the central processor is 8,000 times faster than the card reader of the system. The card reader is by no means slow and processes 1,000 80-column cards per minute.

Dependent upon a variety of factors it takes in the range of 16 man-hours of key punch and key verifier time to keep the card reader busy for one minute, which in turn keeps the central processor busy for .075 seconds. It becomes quite apparent that the input process is quite expensive and still quite slow.

It is to these facts of computer economics that optical character recognition devices address themselves.

Just as the business-type computers had their genesis in the Bureau of the Census of the US Department of Commerce so, too, our experiences with optical scanning for computer input began there. The system known as FOSDIC (Film Optical Sensing Device for Input to Computers) was installed in the Bureau of the Census in 1953 as an ingenious variation of the mark sensing technique developed for punched card processing and, perhaps more familiarly, for test scoring.

Fosdic may be thought of as a machine which is capable of 'reading' information from a microfilm copy of a specifically designed schedule and transferring the data to magnetic tape for processing on a computer. Basically, Fosdic operates as follows: The beam of light locates itself on the microfilm by centring on an index mark (a black square printed on the Fosdic schedule). In the vicinity of each index mark is a series of circles. Marks in these circles convey the information recorded by the person who fills in the document, eg, a Census enumerator. The scanning beam can be controlled accurately enough to move from the centre of an index mark to each of the circles associated with the mark thus enabling it to identify which circles have been marked. A code indicating which circles are filled is recorded on magnetic tape. There are four separate units:

a: The *scanning unit* controls the beam, both horizontally and vertically, and 'decides' where a mark has been made. This unit also electronically adjusts the device to optimize the probability of correctly reading each frame—it measures and adjusts for minor deviations from the nominal 28:1 reduction in photography; it measures and adjusts for tilt of image; and it measures the difference between maximum and minimum light transmission through the frame to establish a range for that frame. This group of steps is called the 'calibration' operation for each frame. Any deviations which cannot be electronically compensated result in a fail calibrate for which a special recording is made on the magnetic tape thus providing the computer with the information that there was a frame of microfilm which was not read.

b: The *program unit* 'tells' the machine what to do. This includes instructions on calibration of the document, on the order in which the questions will be scanned, and on the distances to travel to find the questions and the marking areas. The program unit provides the machine with a great amount of versatility. Documents can be read in a sequence most convenient for succeeding operations. Documents of various sizes can be accommodated. Instructions for the program unit are provided through a plugboard which is wired differently for each different schedule.

c: The *tape unit* holds the magnetic tape, drives it at a constant rate, records impulses on tape, and at the end of a run, rewinds the tape for computer use.

d: The *console* contains all the operating switches, recording dials, cathode ray tube, light beam, film drive mechanism, and an oscilloscope for testing the strength of electrical impulses.

The use of magnetic inks for data processing purposes has never progressed very far in the US Government, its use being largely restricted to commercial banking applications. From Fosdic to the present the use of optical character recognition has been the major emphasis.

The use of highly specialized fonts as exemplified by the Farrington scanner has been limited primarily to systems that are self-contained within an agency. The United States Department of Agriculture is an example of use of the Farrington process. This process is limited to the recognition of unique characters produced by specially equipped typewriters. It costs approximately $40 per instrument to convert to Farrington characters. In essence this process is restricted to internal processing of documents which originate and end in the using organization. The psychological factor of 'difference' has been a deterrent to broader usage for alphabetic data.

Since most major systems in the US Government that use data in such quantities that the initial heavy capital cost of a scanner can be amortized are dependent upon inputs beyond the direct administrative control of the agency, eg, the American taxpayer,

FORM 941
(Rev. Oct. 1964)

EMPLOYER'S QUARTERLY FEDERAL TAX RETURN

U.S. Treasury Department—Internal Revenue Service

FEDERAL INCOME TAX WITHHELD FROM WAGES	1. AMOUNT OF INCOME TAX WITHHELD (If not required write "None")		
	2. ADJUSTMENT FOR PRECEDING QUARTERS OF CALENDAR YEAR		
	3 ADJUSTED TOTAL OF INCOME TAX WITHHELD →		
FEDERAL INSURANCE CONTRIBUTIONS ACT TAXES	4. TOTAL TAXABLE WAGES PAID (From Item 21, Schedule A)		
	5. 7¼% OF WAGES IN ITEM 4		
	6. ADJUSTMENTS (See instructions)		
	7. ADJUSTED TOTAL OF F.I.C.A. TAXES →		
TOTALS	8. TOTAL TAXES (Item 3 plus Item 7)		
	9. TOTAL OF ENCLOSED DEPOSITARY RECEIPTS (From Schedule B, other side)		
	10. BALANCE DUE (Item 8 minus Item 9) PAY TO "INTERNAL REVENUE SERVICE"		

Complete items 11-13 on reverse. If not liable for returns in succeeding quarters write "FINAL" here.

Under penalties of perjury, I declare that I have examined this return, including accompanying schedules and statements, and to the best of my knowledge and belief it is true, correct, and complete.

Date _____ Signature _____ Title _____ *(Owner, President, Partner, Member, etc.)*

BE SURE TO ENCLOSE REMITTANCE, DEPOSITARY RECEIPTS, AND SCHEDULE A WITH THIS RETURN.

T
P
D
I
T

Employer's name, address, identification number and calendar quarter. (If not correctly printed please change)

Name (as distinguished from trade name) — Date quarter ended

Trade name, if any — Identification No.

Address

If form is not preaddressed, check type of employer—
☐ Sole owner ☐ Partnership
☐ Corporation ☐ Other (specify)

---------- Entries must be made both above and below this line ----------

Name (as distinguished from trade name) — Date quarter ended

Trade name, if any — Identification No.

Address

U.S. TREASURY DEPARTMENT
District Director of Internal Revenue

OFFICIAL BUSINESS
POSTAGE AND FEES PAID

POSTMASTER: If undeliverable treat in accordance with Section 355.56 of Postal Manual

SCHEDULE A--QUARTERLY REPORT OF WAGES TAXABLE UNDER THE FEDERAL INSURANCE CONTRIBUTIONS ACT (FOR SOCIAL SECURITY)

IF WAGES WERE NOT TAXABLE UNDER THE F.I.C.A. MAKE NO ENTRIES BELOW

14. Number of persons employed (except agricultural and household employees) during pay period containing the 12th of third month in quarter.	15. Total pages of this return, including this page and any pages of Form 941a.	16. Total number of employees listed.

List for each employee, except agricultural employees, the WAGES taxable under the Federal Insurance Contributions Act (for Social Security) which were paid during the quarter. If you pay an employee more than $4,800 in a calendar year, report ONLY THE FIRST $4,800 of such wages.

SAVE TIME AND MONEY—If your report shows each employee's name and number *exactly* as they appear on his Social Security card it will not be necessary to write back to you to ask for the correct information.

17. EMPLOYEE'S SOCIAL SECURITY ACCOUNT NUMBER (If number is unknown, see Circular E) 000 00 0000	18. NAME OF EMPLOYEE (Please type or print)	19. TAXABLE F.I.C.A. WAGES Paid to Employee in Quarter (Before deductions) Dollars	Cents	20. See instructions on page 2

If you need more space for listing employees, use Schedule A continuation sheets, Form 941a.

Total wages reported in column 19 on this page →

21. TOTAL WAGES TAXABLE UNDER F.I.C.A. PAID DURING QUARTER
(Total wages in column 19 of this page and continuation sheets) $ _____ { Enter this total in Item 4 above.

FILE THIS RETURN WITH YOUR DISTRICT DIRECTOR OF INTERNAL REVENUE.

(Form 941) Page 1

it follows that there are severe limitations to employing a highly specialized typeface.

The Social Security Administration of my Department is a prime example of a major data processing entity that of necessity turned to optical scanning techniques of broad application.

The US social security system is a contributory insurance system under which persons engaged in covered employment and self-employment pay tax on their wages or income up to statutory limits. The system pays benefits to survivors in the event of the death of the covered person, disability benefits in the event of disability and retirement benefits at retirement. The amount of benefit is directly related to the average earnings of the covered person over the course of his employment or self-employment. It is thus necessary to keep accounts on the amount each person earns.

In this largest insurance system in the world some 140 million separate accounts are maintained. The largest single source of input to this file is the tax returns filed by the 3.4 million employers of the United States representing 56.7 million separate accounts on 6.5 million pieces of paper. This vast assemblage of paper recurs each calendar quarter.

Some 425 key punch operators are employed to convert these mountains of paper to additional mountains of punched cards and thence to closely packed reels of magnetic tape.

Since the inputs to this vast system come from essentially every business organization and every self-employed person in the United States you would expect that the writing devices used cover the spectrum. And indeed they do.

Those reporting units which employed 100 or more employees were analyzed and it was found that tax returns were prepared as follows:

Handwritten	.27 per cent
Typewriter	30.70 per cent
Posting-Book-keeping Machine	9.73 per cent
Addressograph	10.09 per cent
Business Machines	33.86 per cent
Combinations	15.35 per cent

On the basis of this data a request for a proposal was prepared so that the industry could suggest equipment that would master this workload in an economic and timely manner. A total of seven responses were received ranging from volunteers to perform research and development on a fully funded basis to proposals with contractual guarantees to perform the work in question so as to amortize the capital investment in two years or less.

The system chosen for this function is the IBM 1975 Model 3 Optical Page Reader which employs as its logic unit the IBM 360 Model 30 Computer. This device is now undergoing acceptance tests in the headquarters of our Social Security Administration.

IBM optical character recognition system

The 1975 Model 3 Optical Page Reader has an effective speed of 64,800 input lines per hour. As input the machine will accept on specially printed forms all data except those which are: handwritten—prepared in blue inks—prepared on proportionally spacing machines—prepared by Addressograph—prepared on certain typewriters with unusual spacing.

Approximately 50 per cent of the total load is suitable for scanning. The scanner will be able to do the work of approximately 200 key punch operators in approximately 2,000 man-hours and relieve computers used to convert cards to magnetic tape of about 885 hours of work per year. Acquisition cost of the 1975 Model 3 is $775,000 plus $148,585 for the 360-30. Operating costs inclusive of maintenance are approximately $30,000 per year.

Since the equipment will displace costs of $1.5 million per year, amortization will occur in 7.6 months exclusive of the savings to be achieved in the use of floor space. Working backwards one can determine the level of operation needed to make the use of the scanner in question for the type of operation described feasible. Rough calculations indicate that from 35 to 40 key punch positions would have to be displaced if the necessary capital investment is to be written off over 5 years.

Performance characteristics

Maximum horizontal beam speed while reading is 138.5″ per second.

Minimum character spacing is 12 characters per inch (.083″ per character).

Transport time (including document registration, indexing of the document, the gap between documents, and document movement from the scanning station) is approximately 197 milliseconds for Form 941 (the face sheet of the tax return), or 256 milliseconds for Form 941A (the continuation page of the tax return which has more lines to be scanned than the face sheet), when these forms are read completely.

Beam start-up, reading the document code, and setting up on the first data line requires approximately 38 milliseconds.

Beam setup on each succeeding data line requires approximately 14 milliseconds.

Scanning a data line 6.3″ long requires approximately 27 milliseconds.

End of document recognition and page total reading require approximately 53 milliseconds.

The US Post Office optical reader program

The US Post Office Department processed 69,676 million pieces of mail in the fiscal year 1964. This is approximately one piece per person per day. To handle this workload, 585,313 employees were on the payroll at the end of the 1964 fiscal year. It is estimated that the mail volume will exceed 100,000 million pieces before 1975.

To keep pace with this ever-increasing volume, the Post Office has been engaged in an automatic address-reader research program since 1954. During this period, they have increased the recognition capability of the machine from its ability to read 'DC' only to a full city/state plus ZIP Code capability, and are considering recognition of the street numbers to enable sortation to carrier scheme.

Mail handling processes
The basic goal of the address reader program is to read and sort machine-addressed and letter-size mail when coupled to a letter sorting machine.

Letter-size mail is defined as mail pieces $3'' \times 4\frac{1}{4}''$ to $5\frac{3}{4}'' \times 11\frac{3}{4}''$ with thickness of .007″ (postal card) to .25″. At present, this size mail is sorted either manually or by the use of letter sorting machines. Letter sorting machines are used in large offices to extend the range of sorts on one operation from 49 to 279. Their method requires the postal operator to sort according to a scheme which is memorized.

Unique requirements
In most industries, the product is defined. Thus, handling procedures are established, rates and performance predicted, etc., and the process is established.

Although the co-operation with large mailers is excellent, the POD does not specify address format, font, character density, or any other factors influencing readability of the mail. They have, however, conducted statistical studies both within the department and on an outside expert basis and established certain criteria. In addition, live mail has been processed in their laboratory, which provides confirmation of statistical work.

Engineering studies, statistical analysis, as well as outside contracts have established many important characteristics of mail pieces. For example, it has been determined that over 98 per cent of the time, the lowest line of the address is located in an area $2'' \times 8''$, starting $\frac{1}{2}''$ from the bottom edge of the envelope and starting 1″ from the left edge, without regard to its size.

The Post Office has a suggested format for locating ZIP Code. It is recommended that the ZIP Code be located on the city/state line with no less than two-tenths of an inch and no more than six-tenths of an inch between the state name and the ZIP Code beginning.

Address-reader operation
ZIP Code and address readers perform the following basic operations:

Feed and transport. Each mail piece is fed (single file, faced and cancelled) in front of the scanning station.

Video processing. Included in this section are circuits for line find, height find, normalization and other formatting problems. Upon completion of these functions, recognition of the address takes place.

Information processing. Upon satisfactory recognition of the address on the envelope, it is compared with a stored program in the reader.

Letter sorting machine controls. A pocket sort code is associated with each recognizable stored code in the information processing section. At this point, an electro-mechanical device known as a code wheel setter is activated. When code wheels match with code bars the mail piece is deposited into its correct pocket.

Problem area
While character density, pitch, font, and the like cannot be ignored, and most certainly will affect machine recognition, optical readers have other factors to consider in their operation.

Interference. Mail pieces with information for in-house routing such as B5451, can cause errors in processing if this interference is in the area normally containing the ZIP Code. It has been established that three-line addresses are more frequent than two-line addresses, a fact that aids the machine in reducing interference when, for example, alpha-numeric designations appear as AB over PA. Thus, such extraneous information would not be interpreted as a legitimate state abbreviation, but would be rejected.

Character height and pitch. Each reader must accommodate characters .08″ to .20″ tall, and between 8 and 14 in pitch. Random sampling shows that while most printed information lies in the area between .09″ and .11″, it is worthwhile to pursue larger and smaller characters.

Colour. Since these machines are optical devices, contrast is important. Addresses in normal colours, provided reasonable contrast is present, will be recognized. Colour on colour, such as dark blue print on light blue envelopes, usually does not work well.

Three major development areas
The current Post Office Department optical reading research has three major work areas. These areas are *ZIP Code Readers* —Burroughs Corporation, National Cash Register Co., and

Rabinow Electronics Inc.; *Alpha-Numeric Reader*—Philco Corporation; and *Presorter*—FMC Corporation.

ZIP code readers

Burroughs Corporation: This method incorporated a flying spot cathode-ray scanner and a recognition concept designated as the 'nearest neighbour' technique.

An electronically deflected spot of light sweeps the face of the moving envelope and locates the bottom line of the address. The spot is then positioned near the end of the bottom line and the ZIP Code is scanned. An electronic signal patter is generated by light-sensitive photo-tubes which 'see' reflections from the envelope face.

This pattern is processed using computer techniques and is statistically analyzed in resistor matrices where the five-digit ZIP Code is compared with a stored address directory, and instructions are sent to the letter sorting machine.

National Cash Register Company: This technique uses two flying spot scanner tubes and a digital correlation method to accomplish character recognition. Separate cathode-ray tubes are used to line find and read scan. The read scan signals are fed to a storage tube which stores each character sequentially and normalizes the image with respect to height and width. The image in the storage tube is scanned and fed into correlation and decision circuits for comparison by digital techniques to the output of a numeric reference generator. Each numeric character is recognized by computing 'A' correlation factors and analog linear 'B' function values. The five-digit ZIP Code is next compared with the address directory and instructions sent to the letter sorting machine.

Rabinow Electronics Inc.: This technique involves a scanning method whereby reflected light from the moving address is projected on to the end surfaces of vertically stacked glass light tubes, each of which conveys the light to separate photo-electric cells. Pulsed-digital techniques are used to process the electrical outputs of these cells for analysis, recognition and comparison with a magnetic core address directory. Instruction signals are then sent to the letter sorting machine.

Alpha-numeric reader

Philco Corporation: The reader employs a flying spot scanner similar to the picture tube in a television set, but designed for a special purpose. Early in 1963, tests at the contractor's plant using envelopes with machine imprinted addresses proved the feasibility of this technique. In August 1963, the 'engineering model reader' was installed in the Postal Laboratory where an extensive series of tests using 'live' envelope mail was conducted. Analysis utilizing a typewriter to simulate a letter sorting machine showed that approximately 70 per cent of the 'live' mail processed during these tests was correctly identified.

Concurrent with these tests, Philco made an intermediate reader employing all of the knowledge gained from the engineering model and using the most advanced components and materials available. The intermediate reader underwent a 50,000-piece live mail test in the Postal Laboratory in October 1964. The results of these tests were most satisfactory, and the knowledge derived from these and other tests now in progress will be incorporated into the final prototype reader. In 1965, the prototype reader was coupled to a Post Office Department letter sorting machine in actual field operation.

Presorter

FMC Corporation: To achieve greatest possible efficiency from an address reader, there is need for a device which can separate machine readable mail from that mail which cannot be read by an address reader. An example of such non-readable mail would be envelopes addressed in handwriting. FMC has a research and development contract to design and fabricate an electronic presorter.

Basically, operation is as follows: reflected light from mail is scanned by fibre optics coupled to photo-sensors; machine imprinted addresses are indicated by high frequency signals from the photo-sensors, while handwriting yields lower frequencies. Thus 'machine readable' is directed into one stacker and 'non-machine readable' into another. The design goal is to make the readable versus non-readable selection with at least a 95 per cent probability that a proper decision has been made.

The long-range effects of the optical reading program can be summarized as follows:

An improved processing rate will be realized so that the projected mail volume of 100,000 million pieces in or before 1975 can be efficiently handled.

The current human memory requirements will be substantially reduced as the address readers will assume the burden of storage and retrieval of the necessary information to direct the mail piece to its required destination with a minimum of error and delay.

Financial savings will be realized. The nature of the process is such that an inherent economic capability is contained therein. The readable mail is processed on a continuous basis without the need for rehandling by either manual or mechanized methods. While I have not made a definitive study of input speeds by various keying devices, I have done the following:

Conducted cyclometer measures of typing output in terms of keystrokes produced in multiple offices of the Social Security Administration. Against a measure of 83 keystrokes per minute, we found that so-called full-time typists produced less than two hours of typing per eight hour day.

Conducted stop-watch time studies in conjunction with cyclometers to determine the increase in keying productivity of an electric over a manual typewriter. A 66 per cent increase was found to be fairly representative.

Designed a basic system to measure the keystroke productivity of IBM 024 and 026 key punch operators. We measured against the following:

Three classes of source documents, ie good, fair, and poor.

Three classes of basic punching, ie numeric, alpha-numeric, and alphabetic.

Three classes of fields to be key-stroked, ie 1 to 30 columns, 31 to 60 columns, and 61 to 80 columns.

From the above, 27 indices of average performance may be derived. The quality of the source document is an important factor. For example, we achieved an average of 101 cards per hour from good source documents, 57 cards per hour from fair source documents and only 37 cards per hour from poor source documents.

Flexowriters used to produce paper tape input for the Medlars system average 4,000 keystrokes per hour. These data coincide fairly closely with productivity measures established two decades ago by the Metropolitan Life Insurance Company in a study on measuring typing efficiency. While considerable improvement may be expected in basic keyboarding processes it appears from our perspective that the greatest potential for breaking the input 'bind' lies in optical scanning.